DOCUMENT RESUME

ED 150 198 TM 006 939

AUTHOR Hisama, Kay K.; And Others
TITLE A New Direction in Measuring Proficiency in English as a Second Language.
PUB DATE [Apr 77]
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Cloze Procedure; College Students; Comparative Analysis; Efficiency; *English (Second Language); Foreign Students; Higher Education; *Language Proficiency; Language Tests; *Predictive Validity; Scoring; Student Evaluation; *Student Placement; *Test Construction; Testing; Testing Problems; Test Reliability
IDENTIFIERS Comprehensive English Language Test Speakers ESL; Michigan Test of English Language Proficiency; *New Cloze Test (Hisama Lewis and Woehlke); Reading for Understanding Placement Test

ABSTRACT

Since standardized tests of English foreign language seem to lack external validity, a new test, the New Cloze Test (NCT), was developed and validated as a placement test for foreign students. Details of the test design, including paragraph preparation, word deletion, and recommended methods for administration and scoring are discussed. Subjects were 136 students at the Center for English as a Second Language, Southern Illinois University, all recent immigrants, representing twelve languages. One form of the NCT was administered at the beginning of a training session along with three placement tests: the Structure and Listening sections of the Comprehensive English Language Test for Speakers of English as a Second Language, and the Reading for Understanding Placement Test. Another form of the NCT was given at the end of the session along with the Michigan Test of English Language Proficiency. Scores on all of these tests, as well as a teacher evaluation report were subjected to comparative analysis. The results indicated that the NCT is highly reliable, valid, and efficient, with respect to testing time and ease of scoring. The NCT satisfies all conditions of a good test, although a few technical problems should be investigated in further research. (CP)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED150198

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KAY K. HISAMA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

# A New Direction in Measuring Proficiency in English as a Second Language

Kay K. Hisama, Ernest L. Lewis, Paula Woehlke

Southern Illinois University at Carbondale

A Paper presented at the annual meeting at the
American Educational Research Association
New York 1977



TM006 939

Abstract

A modified cloze procedure was used to devise two forms of a test of English as a foreigh language (EFL). The resulting New Cloze Test (NCT) is based on sound linguistic and psychometric principles. The test was administered to 136 non-native speakers, and the results were evaluated in terms of three major criteria: reliability, validity, and practicality. The NCT was found to be highly reliable (parallel-form, split-half, and KR #20),Valid (predictive & concurrent) and efficient (testing time, ease in scoring). Suggestions for further research were discussed.

The need for a test of English as a foreign language (EFL) proficiency has paralleled a phenomenal rise in international communication. Several tests of EFL proficiency have been developed during the past decade or so. The two most widely used tests are the Test of English as a Foreign Language (TOEFL) and the Michigan Test of English Language Proficiency (MTELP) or the Michigan Test. A recent addition is the Comprehensive English Language Test for Speakers of English as a Second Language or CELT.

However, developments of EFL tests "Preceded under the impetus of necessity rather than under the guidance of a coherent theoretical understanding of the nature of language and communication (Jakobovits, 1970, p. 150)." Therefore, it is not surprising that these tests have been subjected to an increased criticism on both theoretical and empirical grounds.

Theoretical issues of language testing revolve around identifying the most important aspects of language. The traditional approach to foreigh language learning emphasizes the development of linguistic competence. In this approach, various elements and skills of language are made, and language tests are oriented towards the assessment of these elements and skills. A newer theory of foreigh language differ from the traditional one in that it emphasizes the functional aspect of language. The primary function of language is in its use for communication.

Empirical arguments against the assumptions which underlie the TOEFL and other standardized tests have arisen recently. The substantial intercorrelations between the subtests indicate that (1) there is an obvious overlap in format from test to test, or that (2) many common behaviors are required by the various subtests in the battery. For example, the listening test is a test not only of aural comprehension but also of reading skill. because examinees have to read the printed alternatives in which the total word count is as high as 40 in the case of CELT.

The most serious problem of these EFL tests seems to be the lack of external validity, however. Studies showing these tests as predictors of later criterion performance, typically grade point average, produce very low correlations. The fundamental problem appears to be that these tests fail to differentiate the two aspects of validity, the validity of the test as a measure of "English proficiency" and its validity as a predictor of "academic success."

Noting the above limitations of current forms of EFL tests, Domino directs our attention to the practicality or efficiency of the test. "If, indeed, institutions are merely interested in knowing whether a candidate possesses a minimal knowledge of English, then a standard vocabulary test might be as satisfactory (Domino, 1972, p.552)." Domino's remark on the function of English proficiency tests as the assessment of minimal knowledge is a logical consideration. His argument for a simpler test is important if one weighs the unimpressive empirical evidence of these tests against the long testing time and prohibitive expense involved in the development of these tests. His proposal for the use of

vocabulary tests would generate much controversy, though.

Earlier efforts to develop a new type of EFL proficiency test was made by those who were dissatisfied with the existing tests, and who examined the possibility of using the cloze procedure for testing EFL proficiency. The cloze procedure or the cloze test is a generic name given to techniques in which words are deleted from a portion of passages according to some appropriate rules (e.g. every nth words), then asking subjects to fill in these missing words. Various studies have demonstrated that the procedure is not only a reliable measure of English language proficiency of non-native speakers (EFL proficiency), but also correlates highly with other tests of EFL proficiency. However, the cloze procedure in its present form has a number of problems. The solution for these problems is not simple, and requires a thorough investigation of all aspects of test development including test construction, test administration, and processing.

The major purposes of this study were to design a new version of the cloze procedure.(The New Cloze Test or the NCT) in light of recent of theories of language and the principles of educational and psychological measurement, and to validate it as a placement test for a group of foreign students.

## DESIGN OF THE CLOZE TEST

The design of the New Cloze Test encompassed various aspects of test construction, its administration, and scoring. Generating ideas in order to alleviate these and other problems is the heart of the design process.

### Paragraph Preparation

Paragraphs can be adapted from appropriate reading materials, but they should be carefully edited for meaning, vocabulary, and structure. Specifically each paragraph should have a title descriptive of its content. Special attention should be given to the development of each topic so that the examinees can remember given information accurately. The most important consideration was that the content of each paragraph is meaningful to the examinees because it is within their cultural and experiential domain. There were other considerations.

The difficulty in securing unbiased sampling of vocabulary presents additional problems. In general, reliability and validity of vocabulary tests are much lower than those of other subtests. Therefore an effort was made to minimize the effect of vocabulary on the New Cloze Test scores; words deleted were limited to those which are on the Dale List of 3000 Familiar words (Dale, 1948); i.e. difficult words were replaced with synonyms on a lower level.

With regard to the grammatical structure of paragraphs, complex structures which appear only in written forms were avoided. Use of misleading grammatical structures (e.g. double negatives) was also avoided. The length of each sentence was no longer than

that of a typical spoken sentence.

Finally, the desired length of each passage was determined by its oral presentation time, which is appropriate for the short term memory of an average person. This was set to be less than one minute, the most commonly used time span for many T.V. commercials.

Deletion of Words

A minimum of 10 words were deleted from a single paragraph. Each word was replaced by an underlined blank (12-15 spaces long and all spaces were of equal length). The most common approach recommends the deletion of every nth word where n is a deletion ratio. This method can be called the fixed-ratio approach.

In contrast, the approach followed in this study can be called the variable-ratio approach where the intervals vary around n. This change allowed us greater freedom in designing the New Cloze Test. For example, the test was designed so that the total number of words to be deleted was well balanced in terms of function words and content words. The resulting test score should be a good measure of the proficiency of both the syntatical aspects and of the semantic aspects of language because correct responses to content words reflect the understanding of the meaning of paragraphs on the past of examinees.

Test Administration

It is desirable that the obtained measure is a valid index of the functional aspects of a language. The two functional aspects of a language are the receiving of and the giving of some meaningful information. Specifically, the examinees are to be asked to reconstruct the information given prior to the test

by filling in the blanks. The completed paragraph should be acceptable to native speakers in terms of its structure and meaning. Although this test appears to measure the more basic aspect of the two aspects of information processing (receiving), it is actually tapping the ability necessary for giving information as well, because it requires examinees to reconstruct the original paragraphs.

(1) Presenting the information: each paragraph in its intact form should be presented auditorily prior to the testing of it. In light of psychological theories of information processing, it was felt that the examinnes would be able to apply the higher level mechanism (e.g. Analysis-by-synthesis) if the presentation is repeated. The first presentation would help them to orient themselves toward its overall meaning and its grammatical structure while the second presentation would help them to crystalize these aspects of language. This was examined in the pilot study. It was found that repeated presentation of each paragraph helped the subjects of the pilot study because the second presentation clarified the details of the information given in the first.

(2) Timing: The test is a power test—the examinees should have enough time to complete or to attempt all items. Appropriate timing for each paragraph was also examined in the pilot study. In general, 3-4 minutes for easy passages and 5-6 for more difficult ones were appropriate times.

Scoring the New Cloze Test

The dfsired scoring method is matching each response against that of the scoring key. The highest level of automatic scoring requires the use of a computerized answer sheet. For practical

purposes the use of an answer sheet is desirable.

Responses used for the scoring key were obtained from a group of native speakers who represent the group with whom our subjects need to communicate. Through the unique test construction as well as the unique administration process, there were only one-or two-per-item key responses.

## SUBJECTS

A total of 136 non-native students at the Center for English as a Second Language (CESL) at Southern Illinois University at Carbondale served as subjects. Out of eight sessions in a year, the following three sessions were sampled: October, November, and January. Two criteria for the subjects were that the subject had to be a new entering student to the center, and that he-or-she had not been in the United States longer than one month at the time of testing. The test scores of those who met the above criteria were retained for analyses. They represented 12 different languages, Arabic (N=29), Farsi-Persian (N=49), Spanish (N=27), French and one of the African languages (N=10), Japanese (N=7), and the remaining represented by less than seven subjects each. In addition, two groups of native speakers who enrolled in first semester freshman English Composition ($N_1$=19; $N_2$=27) served as criterion groups. Their responses were used in order to construct keys for the two forms of the NCT.

## METHOD

### Test Administration and Scoring

One form of the NCT was administered at the beginning of the three sessions along with the three placement tests. The

other form of the same test was given along with the Michigan Test during the final week of the same training session; only volunteers took the second tests since CESL did not allow us to call all of those who took the first test for the second testing.

Administration procedures for the two forms of the NCT were as follows. First students were asked to listen to specific instructions on the tape. Following the tryout of a short example test, they were asked to listen to the first passage twice without opening the test booklet. Upon completion of the second presentation of the passage, they were instructed to open the test booklet and to start to fill in the blanks. The same procedure was repeated until they finished all four passages.

Since the NCT is a power test, no special timing for each paragraph was set. In general 3-4 minutes for the first two passages and 5-6 for the latter two were sufficient time. A total testing time for the test ranged from 20-30 minutes, depending on the total number of examinees and their levels of proficiency.

In order to score the NCT, two sets of scoring keys were constructed for the two forms of the test. The keys were constructed in the following fashion. Two groups of students (native speakers of English) who enrolled in freshman English composition were asked to take the test. The number of responses to each item was tabulated. Responses which were made by over 10% were retained to be used as key responses. For most items there was only one, usually the exact work deleted, or two such responses.

The scoring was carried out by giving credit if the response matched those specified in the key.

The Variables

The general scheme of analyses was to compare the NCT with other placement tests already being used at CESL in terms of three criteria of a good test: reliability, validity, and practicality. Two types of criterion variables were used for this purpose: Teacher evaluation and the Michigan Test scores. All variables used for the analyses are listed below.

CELT-Structure (CELT-S): The score on this test represents the percent of items answered correctly; the raw score on a total of 75 items was multipled by a constant, 4/3.

CELT-Listening (CELT-L): The score on this test represents the percent of items correct; the raw score on a total of 50 items was multiplied by a constant, 2.

Reading for Understanding (RFU): The score on this test represents the number of items answered correctly; there are 100 items in this test.

Composite (CELT-S/CELT-L/RFU): The composite score of the above three tests; the sum of the three tests was divided by 3.

The New Cloze Test (NCT): The score on this test represents the percent of items answered correctly; the raw score on a total of 49 items (Form A) or 50 items (Form B) was multiplied by an appropriate constant, 100/49 or 100/50, respectively.

The Michigan Test (MGN): Two subtests are Michigan Test of Aural Comprehension (MGN-L) and Michigan Test of English Language Proficiency (MGN - G + V + R). An average of these two subtests

(scaled scores of the maximum 100 points) is the composite of the Michigan Test (MEN-Composite).

Teacher Evaluation (TEV): LEVEL — the recommended level for the following session and LEV-Rank—the LEVEL in which the relative class standing is taken into account. Both variables were defined with regard to three major subject areas: Oral, Reading, and writing.

The Analysis

All statistical analyses were carried out on IBM 370 (model 158) using the following computer programs.

ANLITH , an item analysis program, programmed by Gronek and Tyler (1972), in order to obtain the followin statistics: scores, means, and standard deviations for the total test; score distribution; item-test correlations; a homogeneity analysis (Kuder-Richardson #20), item difficulty, and frequency distribution of the responses to each item.

Subprogram Pearson Corr (Pearson Product-Moment Correlation Coefficients) in Statistical Package for the Social Sciences (SPSS), Second edition (Nie et.al., 1975), in order to obtain correlations among all the variables defined for this study (see The Variables).

Subprogram Breakdown in SPSS (Nie et.al., 1975) in order to obtain means and standard deviations of subgroups.

## RESULTS

Means and Standard Deviations

Table 1 shows the means and standard deviations of predictor variables for students in the different levels. Generally speaking, the means of all tests demonstrate stable differences between the

Table 1

MEAN SCORES AND STANDARD DEVIATIONS FOR DIFFERENT LEVELS OF PROFICIENCY WITH ENGLISH AS A FOREIGN LANGUAGE AT THE CENTER FOR ENGLISH AS A SECOND LANGUAGE (CESL)

| | | Tests ** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LEVELS* | N | CELT-S | | CELT-L | | RFU | | Composite | | New Cloze | |
| | | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 1 | 2 | 8.00 | 11.31 | 16.00 | 22.63 | 11.50 | 16.26 | 11.83 | 16.73 | 0.0 | 0.0 |
| 11 | 28 | 20.33 | 7.86 | 22.93 | 10.96 | 26.18 | 11.29 | 23.15 | 7.86 | 12.17 | 12.14 |
| 111 | 35 | 32.46 | 9.71 | 32.23 | 9.24 | 40.23 | 9.39 | 34.97 | 5.39 | 39.30 | 12.14 |
| 1V | 36 | 48.96 | 13.54 | 40.11 | 11.07 | 49.72 | 11.27 | 46.26 | 8.78 | 56.07 | 15.86 |
| V | 20 | 55.87 | 9.01 | 50.00 | 10.16 | 55.05 | 8.16 | 53.64 | 6.37 | 70.92 | 8.83 |
| V1 | 12 | 74.55 | 7.70 | 65.50 | 6.50 | 64.25 | 4.05 | 68.10 | 3.02 | 78.40 | 7.20 |
| V11 | 3 | 76.44 | 10.00 | 75.33 | 9.87 | 77.00 | 3.46 | 76.26 | 7.10 | 95.52 | 1.19 |
| All Subjects | 136 | 42.09 | 20.06 | 38.66 | 16.85 | 44.54 | 16.36 | 41.76 | 16.14 | 46.85 | 26.21 |

* The average of the LEVELS in the three areas--Oral, Reading, and Writing.

** The abbreviations are as follows: CELT-S: A Comprehensive English Language Test for Speakers of English as a Second Language--Structure/ CELT-L: A Comprehensive English Language Test for Speakers of English as a Second Language--Listening/ RFU: Reading for Understanding Placement Test/ Composite: The composite score of the three tests, CELT-S, CELT-L, and RFU.

different levels. A few exceptions are noted, however. CELT-S[12], for example does not differenciate between the two highest levels, indicating a ceiling effect of the test.

Reliability

Of the four common procedures to estimate the reliability of a measuring device, the following three were investigated: Kuder Richardson formula #20, Split-half reliability, and Parallel form reliability. Table 2 shows the original values of the reliability coefficients as well as their projected values under KR #20. The results show that the reliability of the NCT clearly exceeds those of other tests.

Table 2

KR #20 RELIABILITY COEFFICIENTS FOR FOUR TESTS OF ENGLISH PROFICIENCY BASED ON N=136

| Test | N of items | Testing time(min.) | Reliability | | | |
|---|---|---|---|---|---|---|
| | | | Original length | Adjusted(A) (Nof items) | | Adjusted(B)* (testing time) |
| | | | | N=100 | N=49 | |
| NCT | 49 | 25 | .96 | .98 | .96 | .96 |
| CELT-S | 75 | 45 | .94 | .95 | .91 | .87 |
| CELT-L | 50 | 40 | .86 | .93 | .86 | .79 |
| RFU | 100 | 75 | .94 | .94 | .88 | .84 |

Two additional types of reliability coefficients were computed for the NCT. The results are summarized in Table 3.

Table 3

VARIOUS RELIABILITY COEFFICIENTS FOR THE NEW CLOZE TEST

| Split-Half (A) N=136 | Split-Half (B) N=136 | Parallel-Form N=54 |
|---|---|---|
| .96 | .94 | .84 |

Of the two methods of obtaining split-half reliability, Method A is considered to be more appropriate than Method B, because the difficulty level of the two halves is well balanced. It is interesting to note that the reliability based on Method A is identical to that of KR #20. The parallel form reliability of .84 is substantial, considering that only volunteers, mostly from two middle levels, took the two forms of the NCT. Group variability is the major factor in increasing the magnitude of reliability coefficient.

In addition, the rater reliability was estimated by correlating the ratings for one subject area, Oral, because there were at least two instructors who evaluated each student for this skill. The rater reliability is extremely high (r=.98 for LEVEL), indicating a nearly perfect agreement among different instructors in rating their students.

Validity

The statistical definition of empirical validity is simply the correlation of the test scores with a special kind of measure called the criterion. Although there are two distinct approaches (predictive and concurrent validity),this study focused on the predictive validity in accordance with its purpose. Table 4 shows the correlations among two criterion variables based on Teacher Evaluation, and four sets of test scores.

Table 4

CORRELATIONS AMONG CRITERION MEASURES AND FOUR TESTS SCORES

| Criterion* | CELT-S | CELT-L | RFU | Composite** | New Cloze |
|---|---|---|---|---|---|
| Level-Oral | .853 | .796 | .796 | .898 | .857 |
| Level-Reading | .843 | .799 | .787 | .892 | .859 |
| Level-Writing | .843 | .799 | .787 | .892 | .859 |
| Composite LEVEL | .802 | .806 | .871 | .859 | .871 |
| LEV-Rank Oral | .864 | .772 | .801 | .896 | .851 |
| LEV-Rank Reading | .826 | .767 | .798 | .878 | .862 |
| LEV-Rank writing | .854 | .766 | .809 | .893 | .877 |
| Composite LEV-R | .858 | .779 | .812 | .900 | .879 |

*Two types of criterion measure, LEVEL and LEV-Rank, were defined on the basis of CESL instructors' evaluation of their students at the end of one session.
LEVEL--the recommended level for the following session;
LEV-Rank--the LEVEL in which the relative class standing is taken into account. Both variables were defined with regard to three major subject areas: Oral, Reading, and Writing.

** Composite of the three placement tests: CELT-S, CELT-L, and RFU

The highest correlations were obtained between the criterion measures and the composite of the three placement tests. This is understandable because the students are assigned to each level on the basis of this composite, and criterion measures of LEVEL and LEVEL-Rank are highly dependent on this initial assignment. However, the correlation coefficients between the criterion measures and the NCT are almost identical to those between the criterion measures and the composite. This high validity of the NCT is remarkable considering the NCT is not one of the tests which make up the composite score.

Table 5

CORRELATIONS AMONG THE MICHIGAN TEST AND FOUR TESTS BASED ON N=54

| MGN* | CELT-S | CELT-L | RFU | New Cloze |
|---|---|---|---|---|
| MGN-L | .506 | .621 | .597 | .664 |
| MGN (G+V+R) | .700 | .486 | .729 | .703 |
| MGN-Composite | .707 | .654 | .779 | .804 |

* The abbreviations are as follows:
MGN-L, Michigan Test of Aural Comprehension
MGN (G+V+R), Michigan Test of English Language Proficiency
MGN-Composite, the sum of the above two subtests

The Michigan Test: Table 5 shows the correlations among the Michigan Test (MGN-L and MGN (G+V+R+)), the three placement tests and the NCT. The correlations between the NCT and the Michigan Test (Composite) are higher than those of any other test with the exception of one case--the correlation between MGN (V+G+R) and RFU. The correlation coefficient between the NCT and MGN - composite, r=.804, is satisfactory if one considers that less than half of the original subjects (volunteers) took both tests, which in turn reduced group variability.

Practicality

The testing time of the NCT was shorter than that for any other tests, i.e., 1/2 to 1/3 of those for other tests (see Table 2). The ease as well as the time involved for scoring the NCT was comparable to that of the other multiple choice tests. The present form of the NCT does not use separate answer sheets. The decision not to use an answer sheet was made because the testing time for this experimental test was severly limited. It is desirable that the final form of the NCT have a separate answer sheet.

There were two problems encountered in the administration of the NCT. It was easier for the students to look at other students' answers for the NCT than it was for the multiple choice tests. When there were a large number of students, it was difficult to control this problem. Second, the students were eager to look at the printed passage while they listened to it, although they were instructed not to do so. These technical problems need further investigation.

In summary, the NCT satisfies all the conditions of a good test although it has a few problems which must be investigated in further research. Extremely high coefficients of reliability and validity demonstrated by an extremely short form of the NCT might partly be explained by the fact that the difficulty level of the test is tailored for the subjects. However, the major reason for this high efficiency of the NCT is that the test was carefully designed in two important aspects of test development: test construction and its administration. The NCT avoided the multiple choice format, and attempted to measure some fundamental aspects of language which take place in the natural process of communication.

The immediate research problems of the NCT are the development of a separate answer sheet and the finding of solutions for some problems encountered in its administration. These problems will be solved without much difficulty, and when they are, the NCT will probably become a very useful test of EFL proficiency in the future.

REFERENCES

Dale, E. Dale List of 3000 Familiar Words. In Dale, E., & Chall, J.S. A formula for predicting readability, Educational Research Bulletin. 1948, 27, 37-54.

Domino, G. Test of English as a foreigh language: A review in O.K. Buros (ed.), Seventh Mental Measurements Yearbook. Highland Park, New Jersey: The Gryphon Press, 1972.

Gronek, T. & Tyler, T. ANLYH: Item, test and homegeneity analysis, and data scoring program. Southern Illinois University at Carbondale, 1972.

Jakobovits, L... Foreign Language Learning. Rowley, Massachusetts: Newbury House Publishers, 1970.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. Statistical Package for the Social Sciences. (2nd Ed.) New York: McGraw-Hill, 1975.